Article

WIND ENGINEERING

Wind Engineering
2019, Vol. 43(4) 359–376

DOI: 10.1177/0309524X19858251
journals.sagepub.com/home/wie
SAGE

Wind resource clustering based on statistical Weibull characteristics

Chantelle Y Janse van Vuuren and Hendrik J Vermeulen

Abstract

This investigation presents results for the clustering of wind resource data based on statistical Weibull characteristics. The clustering of a chosen geographical area is based on the Weibull and mean wind speed characteristics for each geospatial point for the high energy demand period. The geographically clustered area is chosen from one of the renewable energy development zones, which were identified by the Council of Scientific and Industrial Research. The renewable energy dataset used throughout this study represents the eight renewable energy zones through a meso-scale wind and solar dataset, which spans a 5-year period, at a 15-min temporal resolution. The clustering exercise is aimed at the identification of various geographical areas which best represent a specific independent power producers energy site expectations, while balancing factors such as grid stability and economic and environmental considerations. The study looks into various clustering factors, namely the demand seasons and the energy time of use periods, which correlate to energy production demands for the South African region. The clustering algorithms compared within this study include k-means clustering, the Clustering LARge Applications algorithm, the hierarchical agglomerative algorithm and a model-based clustering algorithm. The initial comparison study yielded the k-means algorithm as the best performing algorithm based on the following internal validation metrics: the Silhouette index, Dunn index and the Calinski-Harabasz index. This clustering method is then subsequently performed on various topical case studies.

Keywords

Clustering, Weibull distribution, k-means, plant siting, renewable energy development zones

Introduction

Optimising grid support by appropriate geographical allocation of renewable energy (RE) capacity is an important consideration for successful implementation of South Africa's long-term RE penetration plan (Seth, et al., 2014). Optimal grid support essentially addresses the challenge of maximising energy yield, while reducing the variability of the residual load, thereby mitigating the operational challenges, such as ramping, associated with RE sources.

The Council of Scientific and Industrial Research (CSIR), as part of phase two of the Strategic Environmental Assessment (SEA) study, identified eight Renewable Energy Development Zones (REDZs) throughout South Africa as preferred areas for the deployment of future wind and solar photovoltaic RE generation. These zones have been selected with the view to support the effective and efficient roll-out of utility size wind and solar development in the future (Council for Scientific and Industrial Research, n.d.). The REDZs span large geographical areas with diverse topographical and climatic characteristics. The available high resolution tempro-spatial wind and solar resource data associated with these areas, due to the size of the underlying database, represents a challenging source to interpret in the context of the siting of generation capacity for optimal grid support. In this context, the application of data reduction methodologies using machine learning techniques, such as clustering, can make a valuable contribution to translate multitudinous RE resource data to meaningful information for high-level interpretation, including optimised plant siting (Milligan & Factor, 2000).

This study investigates the performance of various algorithms, including k-means, Clustering LARge Applications (CLARA) algorithm, the hierarchical agglomerative algorithm and a model-based clustering, for clustering wind speed profiles with the view to develop clustered wind maps. The effects of pre-partitioning of the underlying wind resource dataset along a temporal dimension to accommodate grid support aspects such demand seasons and Time of Use (TOU)

Department of Electrical Engineering, Stellenbosch University, Stellenbosch, South Africa

Corresponding author:
Chantelle Y Janse van Vuuren, Department of Electrical Engineering, Stellenbosch University, 7600, Stellenbosch, South Africa.
Email: jvanvuuren@sun.ac.za

Table 1. Spatial and temporal coverage and resolution of the wind speed data contained in the CSIR dataset (Fraunhofer IWES and The CSIR Energy Centre, 2016).

Parameter	Spatial coverage	Temporal coverage	Spatial resolution	Temporal resolution	Height above ground level
Value	South Africa	2009 to 2013	5 km × 5 km	15 min	50 m, 80 m, 100 m, 150 m

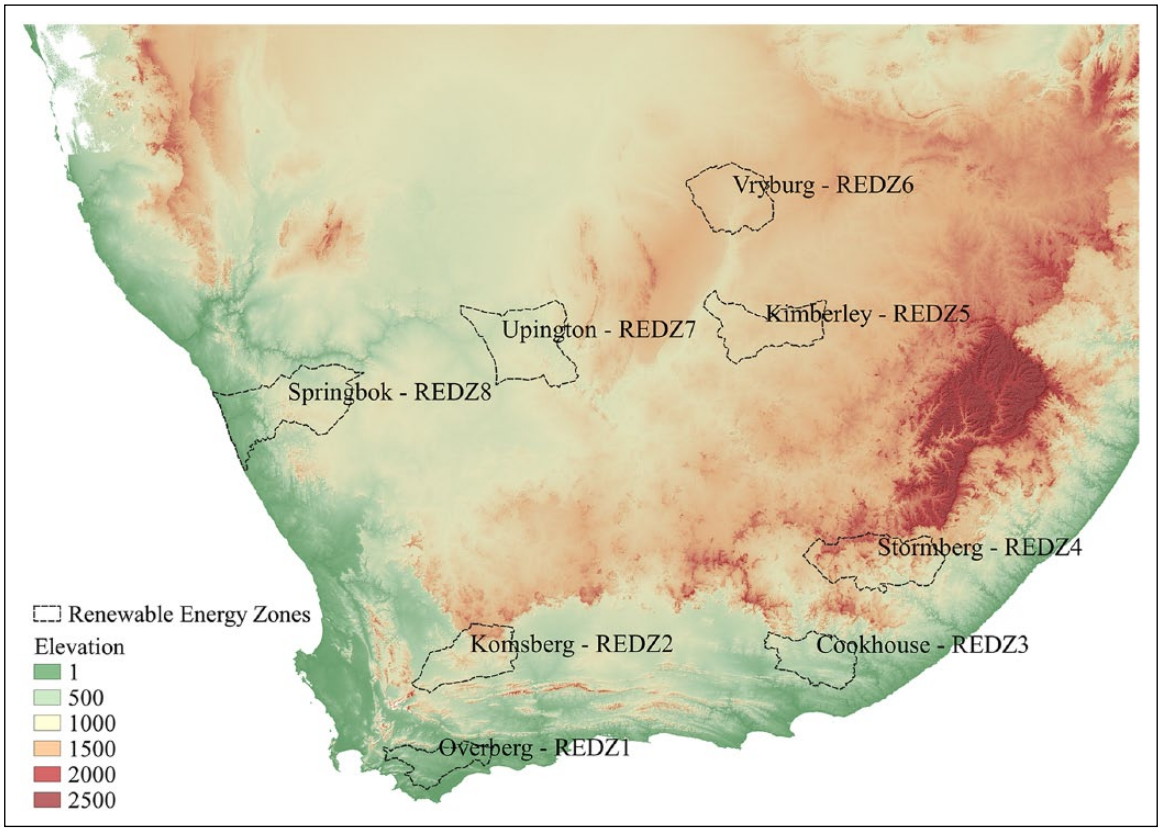

Figure 1. Geographical location of the renewable energy development zones (Janse van Vuuren and Vermeulen, 2019).

considerations are investigated. The clustering algorithms are furthermore applied to the Weibull statistical distributions associated with the tempro-spatial wind resource, rather than the actual tempro-spatial wind speed profiles. These clustered wind maps can assist independent power producers (IPPs) to site according to specific generation goals, which may include highest yield based on demand season, targeted time of use (TOU) periods, etc.

South African RE resource dataset

The temporo-spatial RE resource dataset created by CSIR includes both wind and solar resource data for the Southern Africa region (Fraunhofer IWES and The CSIR Energy Centre, 2016). The dataset has a spatial resolution of 5 km by 5 km and a temporal resolution of 15 min. The wind speed data is provided for various heights, namely, 50, 80, 100 and 150 m above ground level. These dataset characteristics are summarised in Table 1.

The Renewable Energy Development Zones

The REDZs, identified through a SEA study by CSIR, were determined based on criteria such as agriculture, biodiversity, landscape, heritage areas, socioeconomic circumstances and RE yield potential (Council for Scientific and Industrial Research, 2014). The identification of these zones allows for environmental impact procedures to be streamlined and fast-tracked and promotes timeous development of electrical grid support infrastructure. Table 2 lists and summarises the geographical characteristics of the individual zones. Figure 1 shows the geographical locations and elevation characteristics associated with the individual REDZs.

Table 2. Geographical details of the renewable energy development zones (Janse van Vuuren and Vermeulen, 2019).

Zone			Characteristics	
Number	Designation	Province	Area (km^2)	Spatial points within zones
1	Overberg	Western Cape	5263	254
2	Komsberg	Western Cape	8846	341
3	Cookhouse	Eastern Cape	7366	288
4	Stormberg	Eastern Cape	12,041	467
5	Kimberly	Northern Cape	9568	372
6	Vryburg	North West	9204	361
7	Upington	Northern Cape	12,833	497
8	Springbok	Northern Cape	15,214	593

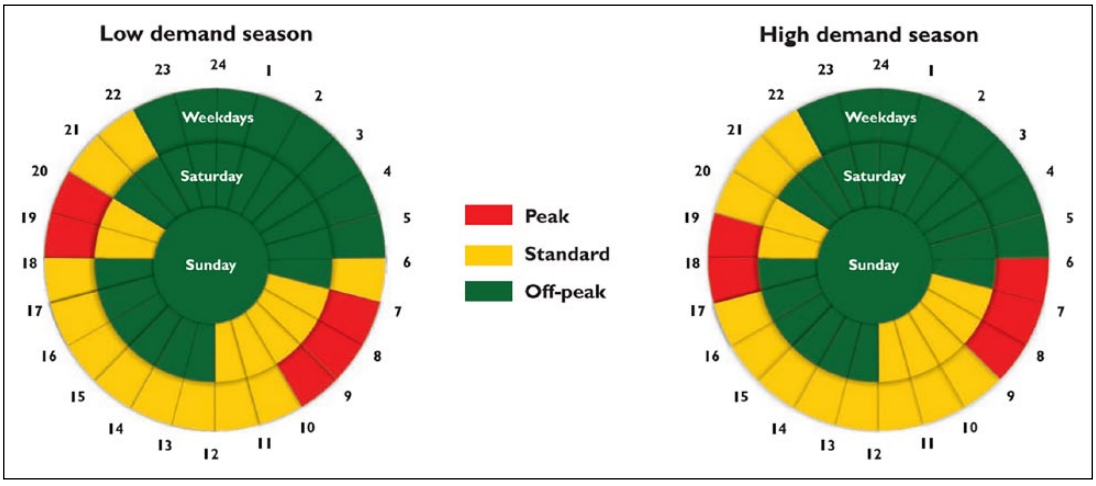

Figure 2. Time of use periods shown for the energy demand seasons (Eskom, 2019).

Table 3. Per unit energy costs of the Megaflex TOU tariff (Eskom, 2019).

Period	Low demand season (1 January to 31 May and 1 September to 31 December)			High demand season (1 June to 31 August)		
	Off-peak	Standard	Peak	Off-peak	Standard	Peak
Cost (pu)	0.14	0.23	0.33	0.17	0.30	1

The South African tariff structure

Eskom, a South African parastatal which accounts for the majority of the countries' power production and distribution, introduced energy TOU tariffs in the early 1990s. The introduction of the TOU tariffs made provision for the high energy demand periods, while hoping to incentivise customers to load outside of peak times, especially in peak seasons, therefore aiding in peak load reduction. A reduction in peak load would, in turn, lower the need for expensive energy generation in energy usage peak periods. Tariff structures are divided into urban and rural areas, where the urban tariffs are available to users who are directly connected to substations. Figure 2 depicts the TOU profiles for the low demand and high demand seasons, where a week's profile has been divided into peak, standard and off-peak representative demand periods. The low demand period, summer, ranges between September and May and the high demand season, winter, from June to August. Table 3 provides the TOU cost in per unit, showing the drastic price increase between demand periods. This stresses the potential for siting RE plants for maximum yield within the higher demand periods, relieving generation demand and decreasing variability of the national energy supply.

Literature review

Overview

The clustering algorithms are applied to the Weibull distributions extracted from the meso-scale, temporo-spatial wind speed profiles for the targeted zones. The Weibull distribution parameters used in the clustering process include the shape parameter, scale parameter, together with the relative mean wind speed. Results are presented for a subset of the temporal wind speed data contained in the CSIR dataset, namely for the high demand season period for 2013, that is, 1 June 2013 to 31 August 2013.

Weibull distribution

Weibull distributions are widely used in practice to represent the statistical properties of wind resource data (Carrillo , et al., 2014). There are two main versions of the Weibull Probability Density Function (PDF), namely the three-parameter and the two-parameter PDFs. The three-parameter PDF is represented by the mathematical equation (Mann, et al., 1974).

The probability density function denotes an element's probability of occurrence based on the elements frequency of occurrence within the chosen dataset. There are two main versions of the Weibull pdf, namely, the three-parameter pdf and the two-parameter pdf. The three-parameter pdf is represented by the mathematical equation (Mann et al., 1974)

$$f(x) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}} \tag{1}$$

subject to

$$f(x) \geq 0; x \geq \gamma; \beta, \alpha > 0; -\infty < \gamma < \infty \tag{2}$$

where β is the shape parameter, β is the scale parameter and γ is the location parameter. When parameters $\gamma = 0$ and $\beta = 1$, in equation (1), this is known as the standard Weibull distribution. In the case where $\gamma = 0$, the equation is simplified to the two-parameter Weibull distribution function. The standard Weibull distribution case reduces equation (1) to the mathematical expression (NIST/SEMATECH, 2013)

$$f(x) = \beta(x)^{\beta-1} e^{-(x)^{\beta}}, x \geq 0; \beta > 0 \tag{3}$$

The Weibull distribution can also be represented in terms of the cumulative distribution function. The cumulative distribution function denotes the probability that the element is equal to or less than a specific instant x, indicated on the x-axis of a cdf graph, with the y-axis denoting the probability of occurrence. The Weibull cdf can be represented by the mathematical function (NIST/SEMATECH, 2013)

$$F(x) = 1 - e^{-(x)^{\beta}}, x \geq 0; \beta > 0 \tag{4}$$

This study clusters based on the two-parameter pdf, together with the relative mean wind speed for each 25 km^2 area block.

Clustering algorithms

The clustering algorithms applied and compared in this study include the k-means algorithm, Clustering LARge Applications (CLARA) algorithm, the hierarchical agglomerative algorithm and a model-based clustering algorithm.

The unclustered dataset, P, can be represented by the expression (Janse van Vuuren and Vermeulen, 2019)

$$P = \left\{p_i, i = 1, 2, 3, \ldots, N^i\right\} \tag{5}$$

where p_i denotes the ith element and N^i denotes the number of observations in the set. The set of clusters, C, can be represented by the expression

$$C = \left\{ C_j \middle| C_j \subset P,\ j = 1,\ 2,\ 3, \ldots,\ N^j \right\} \tag{6}$$

where the set of observations within a cluster, C_j, is represented by the expression

$$C_j = \left\{ C_{jk} \middle| C_{jk} \subset P,\ j = 1,\ 2,\ 3, \ldots,\ N^{jk} \right\}. \tag{7}$$

C_{jk} and N^{jk} denote the kth observation within the jth cluster and the number of observations, respectively, in cluster set C_j. The set of centroids associated with the clusters, W, is represented by the expression

$$W = \left\{ W_j \middle| W_j \in C_j,\ j = 1,\ 2,\ 3, \ldots,\ N^j \right\} \tag{8}$$

where W_j denotes the jth centroid (Janse van Vuuren and Vermeulen, 2019).

In the k-means clustering algorithm, each observation of the set P is iteratively assigned to a cluster C_j with a mean wind speed representative centroid, W_j. The initially appointed cluster assignments of each element, p_i, remains unchanged until convergence occurs. The number of clusters, C_j, for the k-means method must be predefined and methods such as the Elbow method and information criterion approach can be used to determine this predefined number (Sugar and James, 2003; Thorndike, 1953).

The CLARA algorithm is based on the partition around medoids (PAM) method, which has been adapted for use on large datasets (Bhat, 2014). The quality of the selected medoids is determined by the average dissimilarity between each clustered point and its corresponding medoid. This quality measure is identified as the medoid rating function. The set of medoids, M, can be represented by the mathematical set expression

$$M = \left\{ m_j,\ j = 1,\ 2,\ 3, \ldots,\ N^j \right\} \tag{9}$$

where m_j denotes the jth medoid (Janse van Vuuren and Vermeulen, 2019). The rating function, $R(m_j, C_{jk})$, is defined by the mathematical relationship (Bhat, 2014)

$$R\left(m_j, C_{jk}\right) = \sum_{C_{jk}} \frac{d\left(C_{j\mathrm{k}}, \mathrm{rpst}\left(m_j, p_i\right)\right)}{N^{jk}} \tag{10}$$

where $d(C_{j\mathrm{k}}, \mathrm{rpst}(m_j, p_i))$ represents the dissimilarity between two dataset elements C_{jk} and $rpst(m_j, p_i)$ and $rpst(m_j, p_i)$ represents the medoid closest to an element p_i. Using this measure, after suitable repetition, the cluster with the smallest dissimilarity sum is retained.

The agglomerative clustering method is often referred to as a 'bottom-up' approach, whereby each element is initialised as its own cluster. Thereafter, clusters with the smallest sum-of-squares distance between them are merged with successive levelling within the process. This merging process is represented by a dendrogram, which is a tree-like structure that represents the clustered elements within the dataset. This iterative process continues until one large cluster, representing the entire dataset, is reached (Martha et al., 1987).

Model-based clustering, classification and density estimation is based on Gaussian mixture modelling. In the unclustered set of P, the distribution of each element is represented by a probability density function through a finite mixture model of p_i components. The probability density function, $f(p_i, \psi)$, is represented by the mathematical expression (Scrucca et al., 2016)

$$f\left(p_i,\psi\right)=\sum_{k=1}^{P}\pi_k f_k\left(p_i;\theta_k\right) \tag{11}$$

where ψ represents the mixture model parameters, given by

$$\psi=\left\{\pi_1,\ldots,\pi_{P-1},\theta_1,\ldots,\theta_P\right\}. \tag{12}$$

$f_i(p_i;\theta_k)$ denotes the kth component density for the observation p_i, where the parameter vector and θ_k, $(\pi_1,\ldots,\pi_{P-1})$ are the mixing probabilities, subject to

$$\pi_k>0 \tag{13}$$

and

$$\sum_{k=1}^{P}\pi_k=1. \tag{14}$$

The Gaussian finite mixture model is fitted by an expectation–maximisation (EM) algorithm. The EM algorithm is used in conjunction with statistical models, whereby it iteratively finds the maximum possibility of parameters with hidden latent variables (Boyles, 1983)

Validation metrics

The validation methods used to compare the performances of the various clustering algorithms include the Silhouette index, the Dunn index and the Calinski-Harabasz index, together with the adjoining metrics: the average intra-cluster distance and the cluster connectivity.

The Silhouette index is a clustering assignment measure based on the separation and compactness of the formed clusters. This measure lies between the interval [–1, 1], with 1 denoting perfect clustering results and poorly formed clusters observe a silhouette width near −1 (Kim et al., 2017). The silhouette coefficient for the ith element in the dataset in a cluster S_i is defined by the mathematical relationship (Kim et al., 2017; Thinsungnoena et al., 2015)

$$S_i=\frac{\left(b_i-a_i\right)}{max\left(a_i,b_i\right)} \tag{15}$$

where a_i and b_i denote the mean intra-cluster distance and nearest inter-cluster distance, respectively, for each ith element.

The Dunn index is also a measure based on cluster separation and compactness, which ranges between [0:∞] and should be maximised for optimal results. The Dunn index is represented by the mathematical expression (Legány et al., 2006)

$$D=\min_{i=1,\ldots,N^j}\left\{\min_{j=i+1,\ldots,N^j}\left(\frac{diss\left(c_i,c_j\right)}{\max_{m=1,\ldots,N^j}\left(diam\left(c_m\right)\right)}\right)\right\} \tag{16}$$

where

$$diss\left(c_i,c_j\right)=\min_{x\in c_i,y\in c_j}\left\{d\left(x,y\right)\right\}\text{ and }diam\left(c_i\right)=\max_{x,y\in c_i}\left\{d\left(x,y\right)\right\} \tag{17}$$

where N^j denotes the number of clusters, c_i is the ith cluster, $diam(c_m)$ represents the intra-cluster diameter of a cluster m and d(x,y) denotes the distance between two elements x and y.

The Calinski-Harabasz index is a measure of the average inter- and intra-cluster sum of squares, which should be maximised for optimal cluster assignment (Rendón et al., 2011). The Calinski-Harabasz index can be described by the mathematical relationship

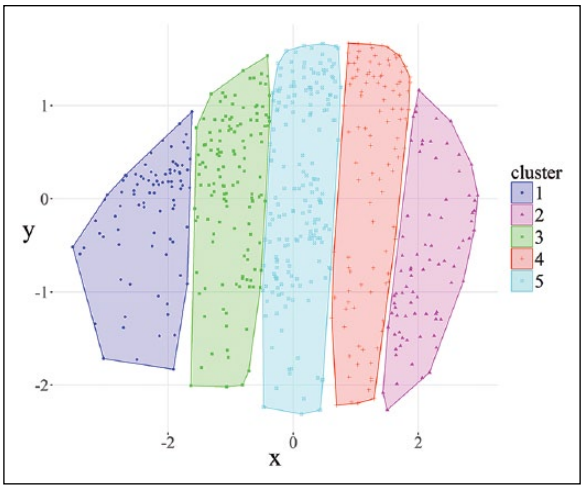

Figure 3. Non-overlapping clusters obtained with the k-means algorithm.

$$CH = \frac{trace(S_B)}{trace(S_w)} \cdot \frac{n_p - 1}{n_p - k} \tag{18}$$

where S_B denotes the inter-cluster scatter matrix, S_w denotes the intra-cluster scatter matrix, n_p is the number of clusters sampled and k denotes the number of clusters.

Case study results

Overview

The case study is performed for the Springbok REDZ using the wind resource data for a single demand season, namely, the high demand season for the 2013 calendar year. The Springbok REDZ is selected for this study as it spans the largest geographical area, therefore having the largest number of associated geographical coordinates. This zone is, furthermore, located on the coast and exhibits diverse topographical characteristics. The high demand season is selected in view of the importance of this period for grid support in the context of seasonal peak demand periods.

A subset of the CSIR dataset is extracted in accordance with the target area and target period. The Weibull distributions and mean values are subsequently derived for each of the geographical coordinates within the target zone.

k-means clustering algorithm

The optimal number of clusters can be determined by the Elbow method (Kodinariya and Makwana, 2013). This method plots the intra-cluster sum of squares measure as a function of the possible number of clusters. In this case, the optimal number of clusters is found as 5, that is, $k = 5$.

Figure 3 displays the non-overlapping clusters formed using the k-means algorithm. The x-axis and y-axis labels in Figure 3 denote a transformed 2D representation of the set of variables obtained using principal component analysis (Jolliffe, 2011). This is a dimensionality reduction algorithm that creates a set of variables representing a projection of the original data set.

Figure 4 shows the Weibull pdf distributions for the averaged wind speed profiles associated with the five clusters, together with the Weibull pdf for the zone as a whole for comparison purposes. This comparison highlights the necessity for further clustering of the zone, as the averaged Weibull pdf distribution masks the zone's high wind speed variability. Figure 5 shows the clustered daily temporal wind speed data, where it can be seen that clustering on the Weibull distribution factors create clear temporal clusters within the daily wind speed profile. Cluster 2 exhibits the highest wind speed, that is, 8.6527 m s^{-1}, while cluster 1 has the lowest average wind speed, that is, 5.5335 m s^{-1}. In the context of TOU periods, this zone displays ideal characteristics, as the temporal profile increases in both the morning and evening peak demand periods. A dip in wind speed is displayed during midday which is ideal when combined with Solar PV energy, as this is the period of highest yield for solar energy production.

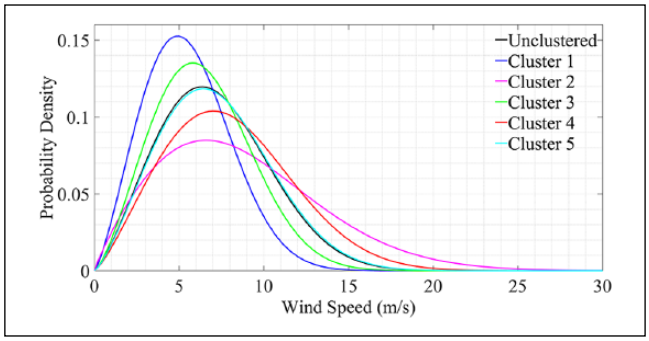

Figure 4. Weibull distributions for the clusters obtained with the k-means algorithm.

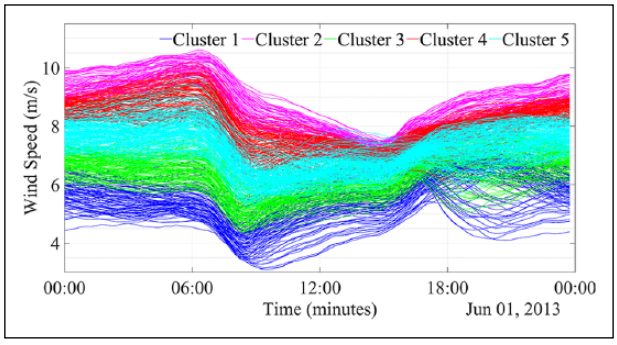

Figure 5. Clustered mean daily wind speed profiles obtained with the k-means algorithm.

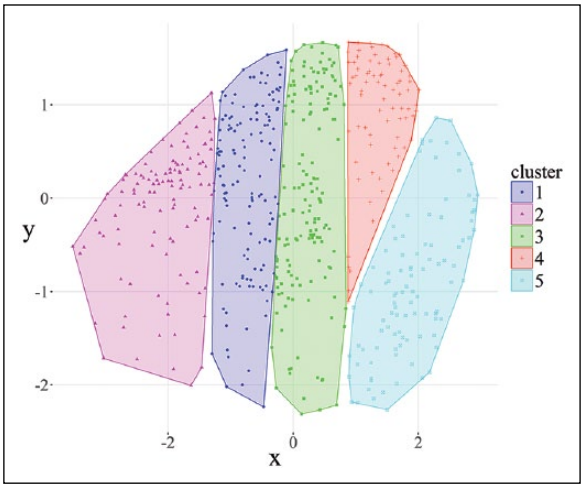

Figure 6. Non-overlapping clusters obtained with the CLARA algorithm.

Clustering LARge applications algorithm

Figure 6 displays the non-overlapping clusters formed when using the CLARA clustering algorithm. Figure 7 shows the Weibull pdf distributions for the averaged wind speed profiles associated with the five clusters, together with the Weibull pdf for the zone as a whole for comparison purposes. This comparison highlights the necessity for further clustering of the zone, since averaging by the entire zone would cause a temporal misrepresentation of the wind speeds at each spatial point. Clustering the larger geographical zone also allows for a decrease in the representative dataset size, which decreases computational time and resources used. Figure 8 shows the clustered daily temporal wind speed data, where it can be seen that clustering on the Weibull distribution factors create clear temporal clusters within the daily wind speed profile. Cluster 5 exhibits the highest wind speed, that is, 8.5409 $m\,s^{-1}$, while cluster 2 has the lowest average wind speed, that is, 5.6776 $m\,s^{-1}$.

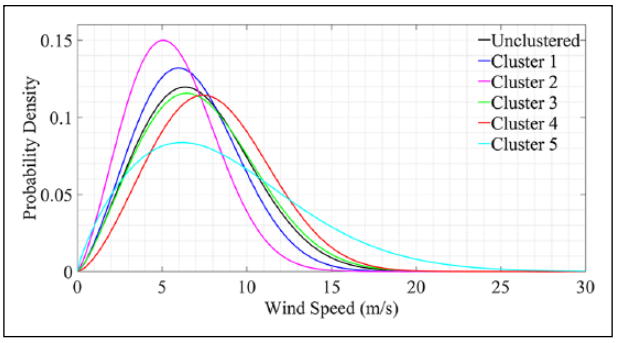

Figure 7. Weibull distributions for the clusters obtained with the CLARA algorithm.

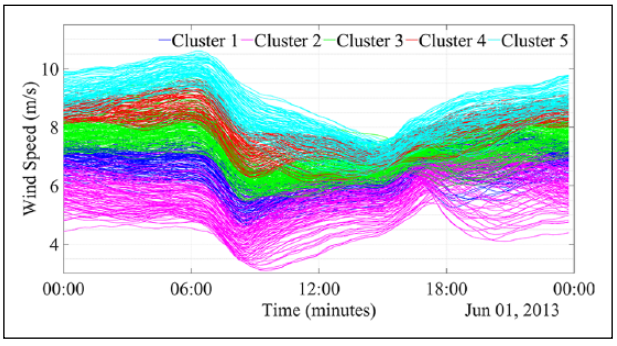

Figure 8. Clustered mean daily wind speed profiles obtained with the CLARA algorithm.

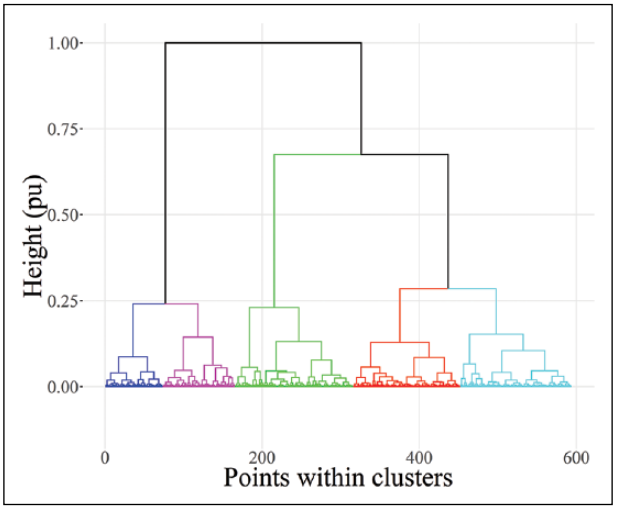

Figure 9. Dendrogram representing the clusters obtained with the agglomerative algorithm.

Agglomerative clustering algorithm

Figure 9 displays a tree-like structure, namely the dendrogram, which displays the clusters formed using the hierarchical agglomerative method on the Weibull pdf distribution characteristics and relative mean wind speed for each spatial point within the Springbok REDZ. The dendrogram shows five clusters, which are relatively equal in size. Figure 10 shows the Weibull pdf distributions for the averaged wind speed profiles associated with the five clusters, together with the Weibull pdf for the entire unclustered zone. Figure 11 shows the clustered daily temporal wind speed data. Cluster 5 exhibits the highest wind speed, that is, 8.5914 $m s^{-1}$, while cluster 1 has the lowest average wind speed, that is, 5.8284 $m s^{-1}$. The k-means,

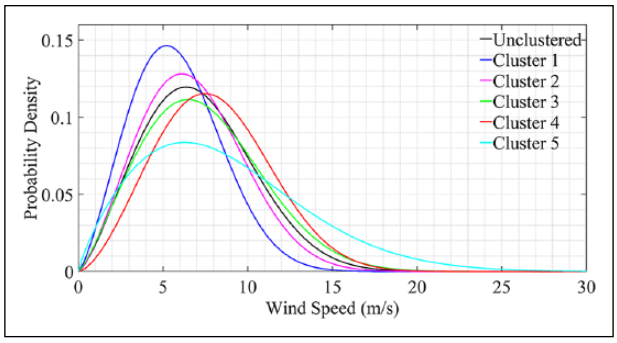

Figure 10. Weibull distributions for the clusters obtained with agglomerative algorithm

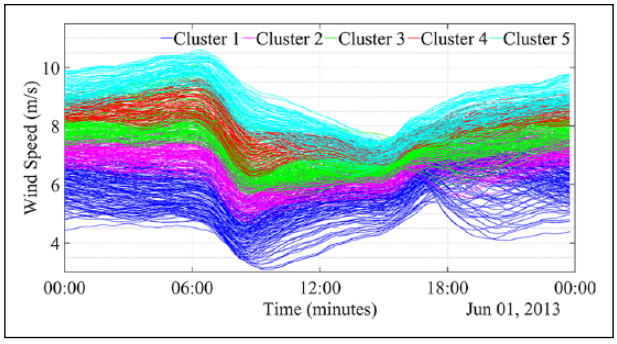

Figure 11. Clustered mean daily wind speed profiles obtained with the agglomerative algorithm.

CLARA and agglomerative clustering methods show similarly clustered temporal outputs, with each method displaying well-defined cluster sections. The cluster outputs for each method do, however, differ slightly in size and mean wind speed.

Model-based clustering algorithm

Model-based clustering is a high-level clustering method. This clustering algorithm uses a Gaussian finite mixture model fitted by an EM algorithm. Figure 12 displays the clusters formed using this method. Figure 13 shows the Weibull pdf distributions for the averaged wind speed profiles associated with the five clusters, together with the Weibull pdf for the unclustered zone as a whole. The Weibull pdf distribution outputs appear less uniformed when compared to the k-means, CLARA and agglomerative clustering methods. Figure 14 shows the clustered daily temporal wind speed data, which does not display clearly segregated clusters. When using this method, clustering on the Weibull and mean wind speed characteristics does not seem to translate into clearly defined temporal wind speed profiles, unlike with the k-means, CLARA and agglomerative methods. These clusters appear to be formed based on the most dominant shape of the daily profiles rather than on a translation of the temporal wind speed positions.

Validation metrics

The clustering methods explored throughout this article are implemented for the selected zone and thereafter require validation in order to identify the most accurate and suitable clustering method tested. This section presents the various validation metrics used to determine the best fit clustering method for the Weibull pdf distribution and mean wind speed characteristics representing the Springbok REDZ. Table 4 presents the various validation metrics and validation characteristics, where the shaded blocks represent the highest scoring clustering algorithm for that validation metric.

It can be seen that the k-means clustering method ranks highest in three of the five validation metrics and the agglomerative method ranks highest in two of the five validation metrics. The CLARA algorithm and the model-based methods underperformed in comparison, with the model-based method performing the worst of the four clustering methods.

It is therefore clear that the k-means algorithm performs most accurately, for this zone, when clustering on Weibull pdf distribution and mean wind speed characteristics. This method displays the maximum silhouette width (which corresponds to the most accurately placed data points within the clusters) and achieved the lowest average intra-cluster distance and the maximum average inter- and intra-cluster sum of squares output.

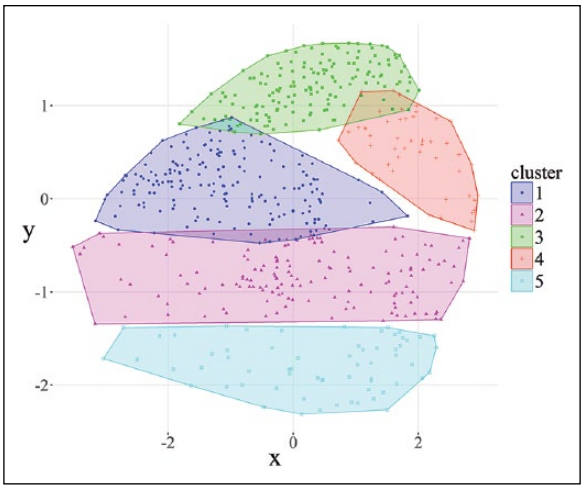

Figure 12. Non-overlapping clusters obtained with the model-based algorithm.

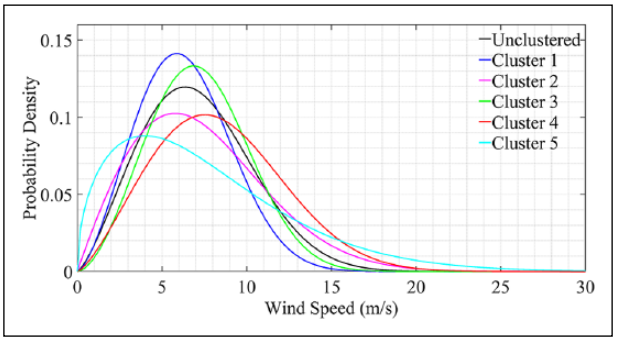

Figure 13. Weibull distributions for the clusters obtained with the model-based algorithm.

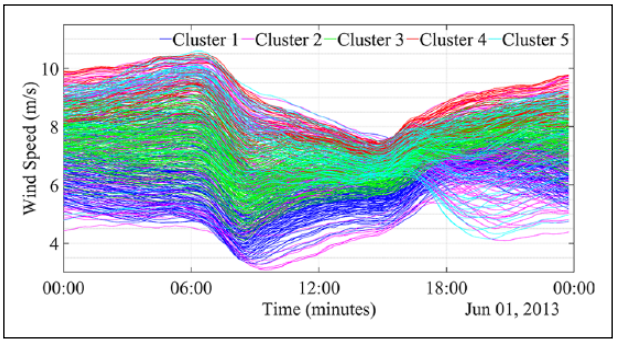

Figure 14. Clustered mean daily wind speed profiles obtained with the model-based algorithm.

Table 4. Validation results for the various clustering algorithms.

Validation method	Agglomerative	k-means	CLARA	Model based
Connectivity	49.1869	69.7282	70.1881	280.9702
Dunn index	0.0278	0.021	0.0089	0.0025
Silhouette index	0.3737	0.3985	0.3931	0.013
Average intra-cluster distance	0.6780739	0.6344	0.6398	1.362557
Calinski-Harabasz	1025.72	1210.686	1140.99	87.80323

CLARA: Clustering LARge Applications.

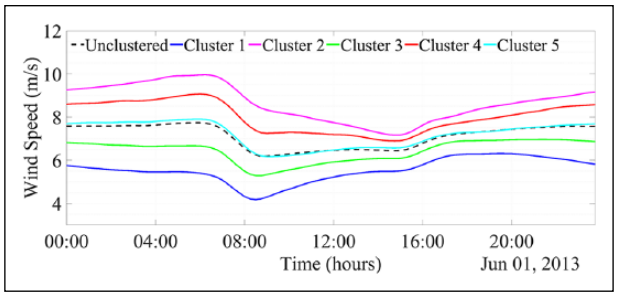

Figure 15. Mean clustered daily wind speed profiles obtained with the k-means algorithm for the high demand season.

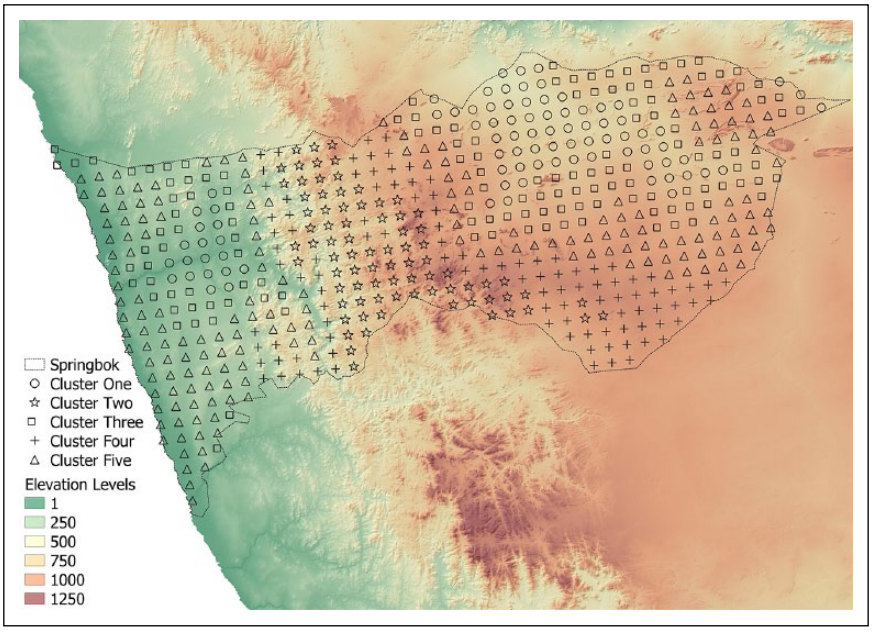

Figure 16. Geographical elevation map of the clusters obtained with the k-means algorithm.

k-means clustering method analysis

As the optimally validated clustering method, the k-means clustering output is further analysed. Figure 15 displays the mean clustered daily wind speed profiles obtained with the k-means algorithm. Figure 16 displays the corresponding cluster outputs on a geographical map. Here, it can be seen that cluster 2, representing the highest wind speed, has corresponding points situated on the geographical areas with the highest elevation levels.

With further analysis of the mean daily temporal wind speed profiles depicted in Figure 5, the results show that clustering on the Weibull distribution parameters and the relative mean wind speeds effectively groups similar temporal wind speed profiles. These cluster parameters are independent of time and location and are based on the entire unaveraged high demand seasons' wind speed characteristics. Despite the temporal-independent inputs, the clustering algorithm managed to correctly cluster on the average daily temporal profile.

As the optimally validated clustering method for the Weibull pdf distribution and mean wind speed characteristics (shape, scale and mean), the k-means method is analysed and compared to clustering on other parameter variations. These variations include shape and scale, shape and mean, scale and mean, as well as the temporal wind speed profiles. Figure 17 displays the various cluster outputs for the parameter variations, where the scale and mean clustering parameters produce well-defined and separated cluster results, similarly found in Figure 15. Figure 18 displays the corresponding temporal cluster outputs, providing a comparison of the optimal parameters to use for temporal wind speed profile clustering. Table 5 provides the internal validation metrics for the varied parameters tested.

From Table 5, it can be seen that the scale and mean parameters produce the highest validation ranking for both connectivity and silhouette width. Clustering on temporal time-dependent characteristics versus the statistical parameters

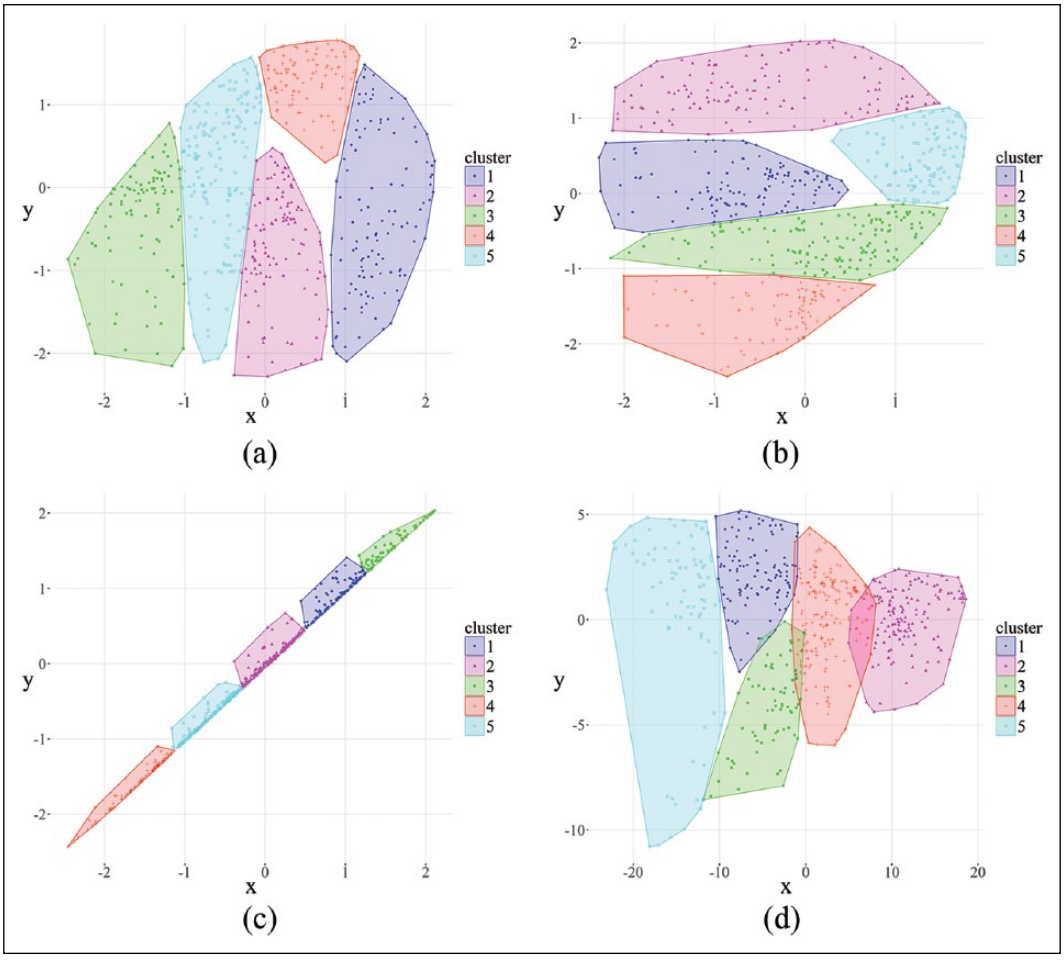

Figure 17. Non-overlapping clusters obtained with the k-means algorithm for varied input parameters: (a) shape and scale, (b) shape and mean, (c) scale and mean and (d) temporal characteristics.

of the wind speed appears to show no significant difference when using the k-means clustering algorithm. When comparing Figure 15 and Figure 18(c), the Weibull shape characteristic displays minimal effect on the clustered output.

Weibull characteristic clustering based on TOU periods

Overview

This section provides a detailed analysis of clustering based on a TOU period for all eight REDZs. The optimally validated k-means method is performed on all eight zones in order to produce clustered outputs which can be selected by IPPs for various siting scenarios. The clustered outputs are then translated into a geographical siting map. The Elbow method is again used to determine the optimal number of clusters, which is shown in Figure 19 as $k=5$.

Clustering based on the peak TOU period

The high demand season is further divided into the peak periods of the weekday demand profile described in the South African tariff structure section. The statistical characteristics are extracted from the peak TOU periods in the high demand season, with the k-means algorithm applied to this targeted subsection. Targeting the peak TOU periods is useful when determining the optimal RE plant placement for high energy yield within the peak demand periods, which would aid in the support of the national energy grid supplier. Figures 20 and 21 show the Weibull pdf and cdf distributions, respectively, for the clustered peak TOU periods. The figures depict an apparent contrast between the five clustered TOU periods, where the probability of occurrence and variability of each cluster differs vastly. These outputs provide insight for IPPs into the statistical nature of the wind resources within each zone, for various periods of interest. The statistical outputs also allow

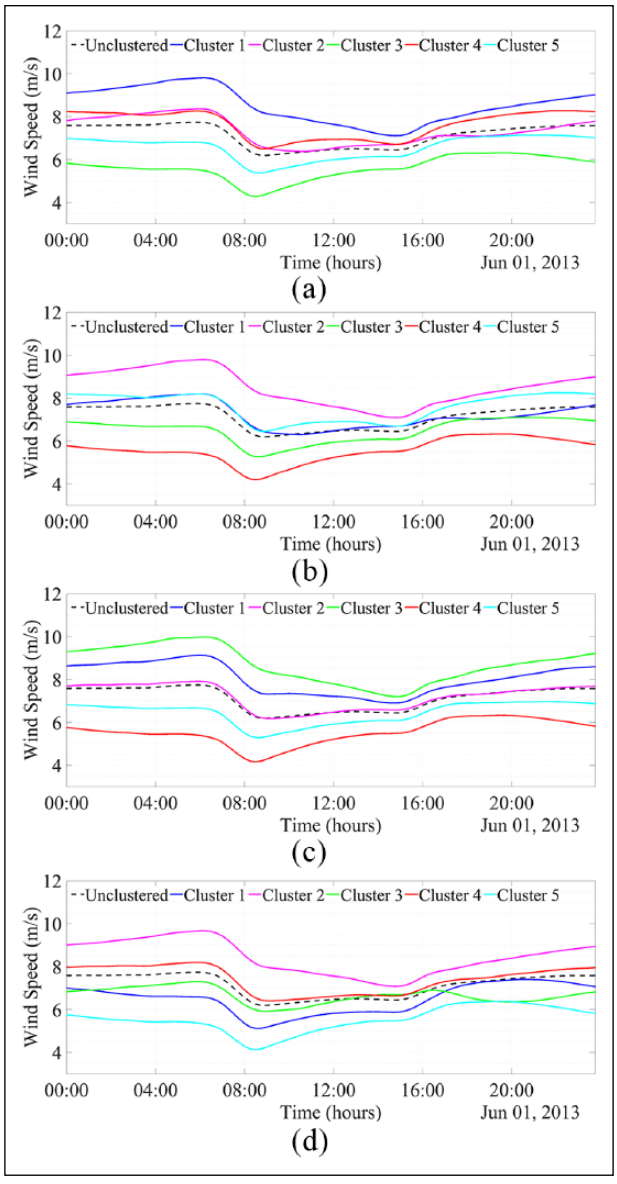

Figure 18. Mean clustered daily wind speed profiles obtained with the k-means algorithm for varied input parameters: (a) shape and scale, (b) shape and mean, (c) scale and mean and (d) temporal characteristics.

Table 5. Validation results for the various clustering input parameters.

Input parameter	Connectivity	Dunn index	Silhouette width
Shape, Scale, Mean	69.7393	0.0179	0.4006
Shape, Scale	68.3532	0.0161	0.4068
Shape, Mean	79.3167	0.0212	0.4106
Scale, Mean	24.0258	0.0101	0.5167
Temporal	105.229	0.0502	0.3845

The shaded blocks depicting the best preforming input parameters for the specific validation metric.

for siting based on a quantifiable weighting of probable yield versus the variability of the resource within the optimal yield period. Figure 22 provides a detailed spread of the wind resource within each cluster, giving the median, lower and upper quartile wind speeds, as well as, the outliers and general wind speed spread.

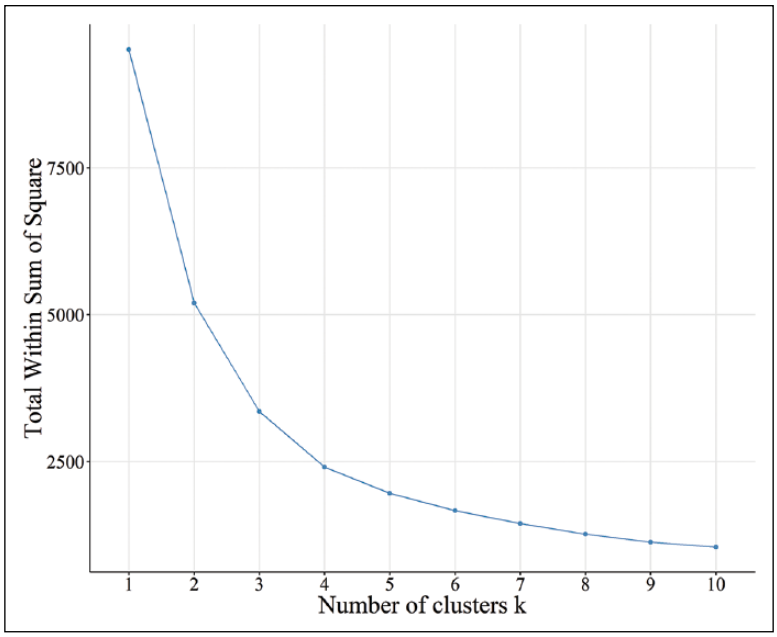

Figure 19. Elbow method depicting the optimal number of clusters based on the k-means algorithm.

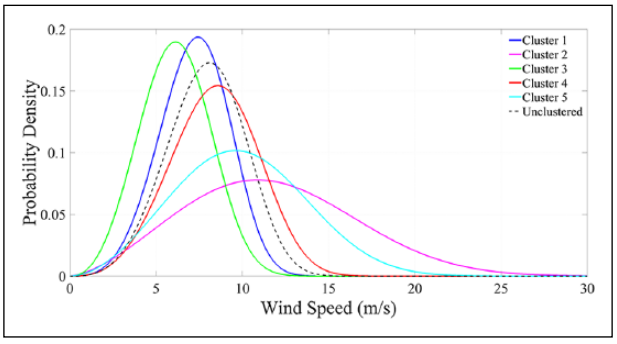

Figure 20. Weibull probability density function shown for the peak period in the high demand season.

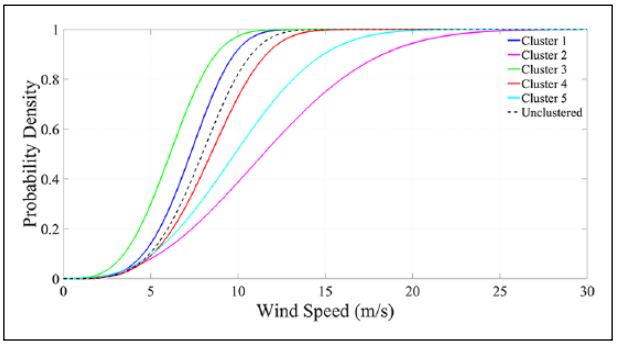

Figure 21. Weibull cumulative distribution function shown for the peak period in the high demand season.

Figure 23 depicts the translated clusters onto a geographical map, where the various statistical and temporal cluster traits are tied to the RE Zone locations. It can be seen that cluster 2, which has the highest mean wind speed and variability, lies predominantly within the middle region of South Africa. Cluster 3, with the lowest mean wind speed for the peak TOU periods, however, is represented mainly in the Northern REDZs. Figure 24 shows a zoomed in map of the clustered

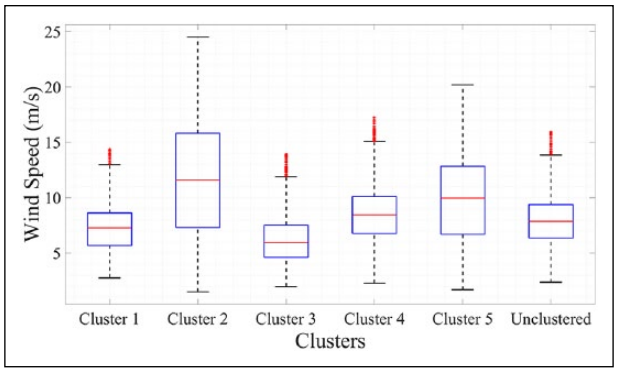

Figure 22. Boxplot showing the 5 representative peak high demand clusters characteristics.

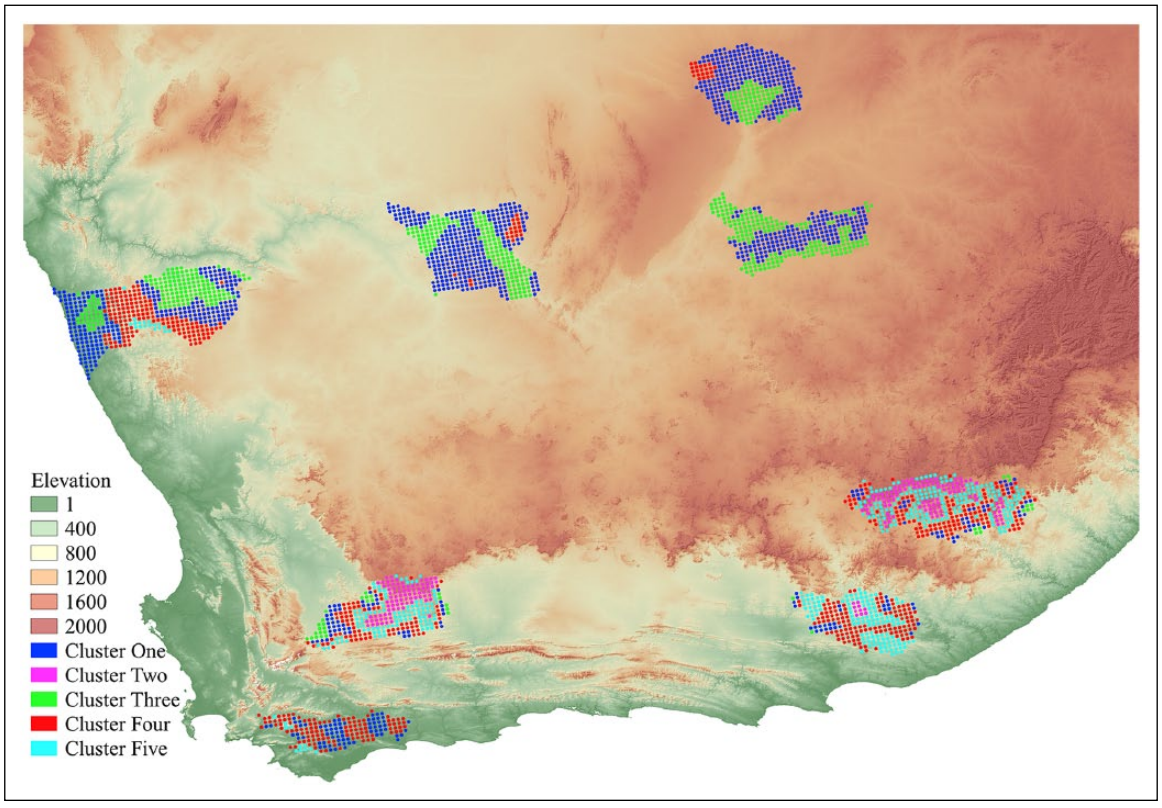

Figure 23. Clustered geographical map representing the peak high demand period.

profiles in the Komsberg Renewable Energy Zone. This clustered map clearly shows the correlation between clustered average wind speed and the topographical features. Cluster 2, represented by the stars in Figure 24, shows the highest average wind speed, which corresponds to the higher elevation levels shown. Contrastingly, cluster 3 has the lowest average wind speed and these clustered spatial points are situated at lower elevation levels.

Conclusion

This study applied various clustering techniques to the characterising Weibull distribution and mean wind speed data for Springbok REDZ. The clustering techniques include k-means clustering, the CLARA algorithm, hierarchical agglomerative

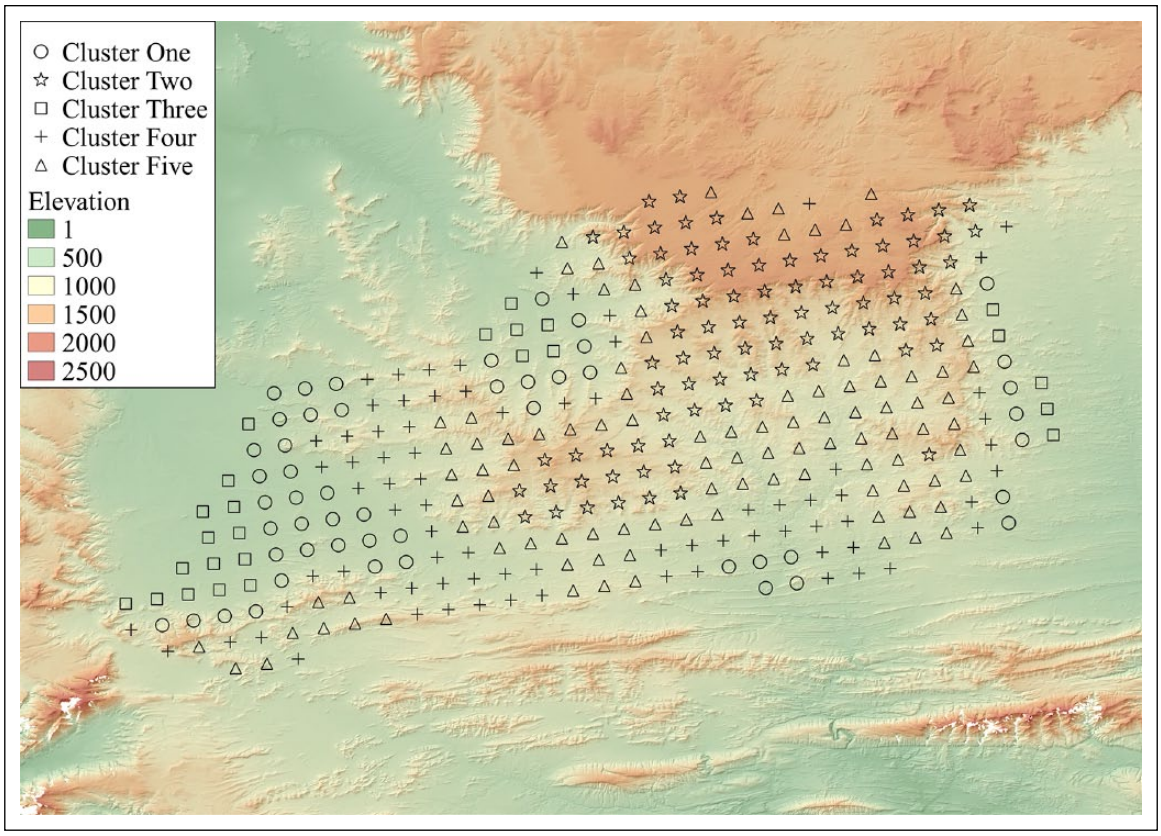

Figure 24. A zoomed in map of clustered profiles in the Komsberg Renewable Energy Zone.

clustering and the model-based clustering method. The outputs of each method were then validated and compared in order to establish the best clustering method on the Weibull distribution and mean wind speed data.

The validation metrics proved the k-means clustering method to be optimal for clustering on Weibull distribution characteristics. This method displays the maximum silhouette width and achieved the lowest average intra-cluster distance and the maximum average inter- and intra-cluster sum of squares output.

The k-means clustering algorithm was then re-run using varied input parameters for comparison purposes. The varied input parameters include shape and scale, shape and mean, scale and mean, as well as the temporal wind speed profiles. The results show that clustering on temporal time-dependent characteristics versus the statistical parameters of the wind speed appears to show no significant difference when using the k-means clustering algorithm.

The k-means clustering method is further applied to peak TOU periods for all eight REDZs within the high demand season. This was done in light of the high energy demand seen on the South African energy grid supply. The aim is to identify areas within the South African region, which will produce a high annual energy yield while potentially decreasing the peak residual energy demand load. These clustered outputs are translated into a geographical peak TOU demand season map, which would visually aid IPPs in identifying the high yield, grid supportive sites within the REDZs.

Future work includes finding the optimal split of wind and solar energy within the defined clusters to best support a decrease in variability on the national energy supply profile.

Acknowledgements

The authors gratefully acknowledge the support of the Doug Banks Renewable Energy Vision, CSIR and ESKOM Tertiary Education Support Program in conducting this research.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors acknowledge the Doug Banks Renewable Energy Vision and ESKOM Tertiary Education for financial support for the research of this article.

References

Bhat A (2014) k-medoids clustering using partitioning around medoids for performing face recognition. *International Journal of Soft Computing, Mathematics and Control* 3(3). Available at: https://www.wireilla.com/ns/maths/Papers/3314ijscmc01.pdf

Boyles R (1983) On the convergence of the EM algorithm. *Journal of the Royal Statistical Society* 45: 47–50.

Carrillo C, Cidrás J, Díaz-Dorado E, et al. (2014) An Approach to Determine the Weibull Parameters for Wind Energy Analysis: The Case of Galicia (Spain). *Energies* 7: 2676–2700.

Council for Scientific and Industrial Research (n.d.) *National Wind Solar Sea*. Available at: https://www.csir.co.za/national-wind-solar-sea (accessed 8 October 2018).

Council for Scientific and Industrial Research (2014) *Strategic Search for SA's Best Wind and Sun*. Pretoria, South Africa: Council for Scientific and Industrial Research.

Eskom (2019) *Tariffs and Charges Booklet*. Johannesburg, South Africa: Eskom.

Fraunhofer IWES and The CSIR Energy Centre (2016) *Wind and Solar PV Resource Aggregation Study for South Africa*. South Africa: Fraunhofer IWES. Available at: https://www.csir.co.za/sites/default/files/Documents/Wind%20and%20Solar%20PV%20Resource%20Aggregation%20Study%20for%20South%20Africa_Final%20report.pdf

Janse van Vuuren CY and Vermeulen HJ (2019) Clustered wind resource domains for the South African renewable energy development zones. In: *Southern African universities power engineering conference/robotics and mechatronics/pattern recognition association of South Africa*, Bloemfontein, South Africa, 28–30 January.

Jolliffe I (2011) Principal component analysis. In: Lovric M (ed.) *International Encyclopedia of Statistical Science*. Berlin; Heidelberg: Springer, 339–341.

Kim B, Kim J and Yi G (2017) Analysis of clustering evaluation considering features of item response data using data mining technique for setting cut-off scores. *Symmetry MDIP* 9(5): 62.

Kodinariya TM and Makwana PR (2013) Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies* 1(6): 90–95.

Legány C, Juhász S and Babos A (2006) Cluster validity measurement techniques. In: *International conference on artificial intelligence, knowledge engineering and data bases*, Madrid, 15–17 February.

Mann NR, Schafer RE and Singpurwalla ND (1974) *Methods for Statistical Analysis of Reliability and Life Data*. New York: John Wiley and Sons.

Martha C, Milligan W and Cooper G (1987) Methodology review: Clustering methods. *Applied Psychological Measurement* 11(4): 329–354.

Milligan MR and Factor T (2000) Optimizing the geographic distribution of wind plants in Iowa for maximum economic benefit and reliability. *Wind Engineering* 24(4): 271–290.

NIST/SEMATECH (2013) Weibull distribution. In: *E-Handbook of Statistical Methods*. Washington, DC: US Department of Commerce, (Section 1.3.6.6.8).

Rendón E, Abundez IM, Gutierrez C, et al. (2011) A comparison of internal and external cluster validation indexes. In: *Proceedings of the applications of mathematics and computer engineering*, Puerto Morelos, Mexico, 29–31 January, pp. 158–163. Stevens Point, WI: WSEAS.

Scrucca L, Fop M, Murphy TB, et al. (2016) mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *NCBI* 8(1): 289–317.

Seth P, Tuler P, Bonnie R, et al. (2014) Wind energy facility siting: Learning from experience and guides for moving forward. *Wind Engineering* 38(2): 203–216.

Sugar CA and James GM (2003) Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association* 98: 750–763.

Thinsungnoena T, Kaoungkuand N, Durongdumronchai P, et al. (2015) The clustering validity with silhouette and sum of squared errors. In: *Proceedings of the 3rd international conference on industrial application engineering 2015*, Kitakyushu, Japan 28–31 March.

Thorndike RL (1953) Who belongs in the family? *Psychometrika* 18(4): 267–276.